PCT/GB 99/03751.

The Patent Office

INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

GB 99/3737

EJU

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed

Dated 22. Nov. 99

CERTIFIED COPY OF PRIORITY DOCUMENT

Patents Form 1/77

Patents Act 19
(Rule 16)

The Patent Office

10NOV98 E403502-1 000085
P01/7700 0.00 - 9824552.5

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

- 9 NOV 1998

The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

THE PATENT OFFICE
L

1. Your reference SP/4819 UK

2. Patent application number
(The Patent Office will fill in this part) **9824552.5**

3. Full name, address and postcode of the or of each applicant *(underline all surnames)*

ROYAL HOLLOWAY
UNIVERSITY OF LONDON
EGHAM
SURREY TW20 0EX
UNITED KINGDOM

Patents ADP number *(if you know it)*

If the applicant is a corporate body, give the country/state of its incorporation UNITED KINGDOM

07547193001

4. Title of the invention

DATA CLASSIFICATION APPARATUS AND METHOD THEREOF

5. Name of your agent *(if you have one)*

"Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*

STEVENS, HEWLETT & PERKINS
1 SERJEANTS' INN
FLEET STREET
LONDON EC4Y 1LL

Patents ADP number *(if you know it)* 1545003

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number

Country	Priority application number *(if you know it)*	Date of filing *(day / month / year)*

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

Number of earlier application	Date of filing *(day / month / year)*

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:*

a) any applicant named in part 3 is not an inventor, or
b) there is an inventor who is not named as an applicant, or
c) any named applicant is a corporate body.
See note (d))

YES

Patents Form 1/77

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form	
Description	13
Claim(s)	4
Abstract	
Drawing(s)	3 + 3

10. If you are also filing any of the following, state how many against each item.

Priority documents	
Translations of priority documents	
Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*	
Request for preliminary examination and search *(Patents Form 9/77)*	
Request for substantive examination *(Patents Form 10/77)*	
Any other documents *(please specify)*	

11. I/We request the grant of a patent on the basis of this application.

Signature Stevens Hewlett & Perkins Date 9.11.1998

STEVENS, HEWLETT & PERKINS

12. Name and daytime telephone number of person to contact in the United Kingdom SARAH PERKINS 0171 936 2499

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) *If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.*

b) *Write your answers in capital letters using black ink or you may type them.*

c) *If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.*

d) *If you have answered 'Yes' Patents Form 7/77 will need to be filed.*

e) *Once you have filled in the form you must remember to sign and date it.*

f) *For details of the fee and ways to pay please contact the Patent Office.*

DATA CLASSIFICATION APPARATUS AND METHOD THEREOF

The present invention relates to data classification apparatus and an automated method of data classification thereof that provides a universal measure of confidence in the predicted classification for any unknown input. Especially, but not exclusively, the present invention is suitable for pattern recognition, e.g. optical character recognition.

In order to automate data classification such as pattern recognition the apparatus, usually in the form of a computer, must be capable of learning from known examples and extrapolating to predict a classification for new unknown examples. Various techniques have been developed over the years to enable computers to perform this function including, inter alia, discriminant analysis, neural networks, genetic algorithms and support vector machines. These techniques usually originate in two fields: machine learning and statistics.

Learning machines developed in the theory of machine learning often perform very well in a wide range of applications without requiring any parametric statistical assumptions about the source of data (unlike traditional statistical techniques); the only assumption made is the iid assumption (the examples are generated from the same probability distribution independently of each other). A new approach to machine learning is described in US5640492, where mathematical optimisation techniques are used for classifying new examples. The advantage of the learning machine described in US5640492 is that it can be used for solving extremely high-dimensional problems which are infeasible for the previously known learning machines.

A typical drawback of such techniques is that the techniques do not provide any measure of confidence in the predicted classification output by the apparatus. A typical user of such data classification apparatus just hopes that the accuracy of the results from previous analyses using benchmark datasets is representative of the results to be obtained from the analysis of future datasets.

Other options for the user who wants to associate a measure of confidence with new unclassified examples include performing experiments on a validation set, using one of the known cross-validation procedures, and applying one of the theoretical results about the future performance of different learning machines given their past performance. None of these confidence estimation procedures though provides any practicable means for assessing the confidence of the predicted classification for an individual new example. Known confidence estimation procedures that address the problem of assessing the confidence of a predicted classification for an individual new example are ad hoc and do not admit interpretation in rigorous terms of mathematical probability theory.

Confidence estimation is a well-studied area of both parametric and non-parametric statistics. In some parts of statistics the goal is classification of future examples rather than of parameters of the model, which is relevant to the need addressed by this invention. In statistics, however, only confidence estimation procedures suitable for low-dimensional problems have been developed. Hence, to date mathematically rigorous confidence assessment has not been employed in high-dimensional data classification.

The present invention provides a new data classification apparatus and method that can cope with high-dimensional classification problems and that provides a universal measure of confidence, valid under the iid assumption, for each individual classification prediction made by the new data classification apparatus and method.

The present invention provides data classification apparatus comprising: an input device for receiving a plurality of training classified examples and at least one unclassified example; a memory for storing the classified and unclassified examples; an output terminal for outputting a predicted classification for the at least one unclassified example; and a processor for identifying the predicted classification of the at least one unclassified example wherein the processor includes: classification allocation means for allocating potential classifications to each unclassified

example and for generating a plurality of classification sets, each classification set containing the plurality of training classified examples and the at least one unclassified example with its allocated potential classification; assay means for determining a strangeness value for each classification set; and a comparative device for selecting a classification set containing the most likely allocated potential classification for at least one unclassified example, whereby the predicted classification output by the output terminal is the most likely allocated potential classification, according to the strangeness values assigned by the assay means.

In the preferred embodiment the processor further includes a strength of prediction monitoring device for determining a confidence value for the predicted classification on the basis of the strangeness value of a set containing the at least one unclassified example with the second most likely allocated potential classification.

With the present invention the conventional data classification technique of induction learning and then deduction for new unknown data vectors is supplanted by a new transduction technique that avoids the need to identify any all encompassing general rule. Thus, with the present invention no multidimensional hyperplane or boundary is identified. The training data vectors are used directly to provide a predicted classification for unknown data vectors. In other words, the training data vectors implicitly drive classification prediction for an unknown data vector.

It is important to note that with the present invention the measure of confidence is valid under the general iid assumption and the present invention is able to provide measures of confidence for even very high dimensional problems.

Furthermore, with the present invention more than one unknown data vector can be classified and a measure of confidence generated simultaneously.

In a further aspect the present invention provides data classification apparatus comprising: an input device for receiving a plurality of training classified examples and at least one unclassified example; a memory for

storing the classified and unclassified examples; stored programs including an example classification program; an output terminal for outputting a predicted classification for the at least one unclassified example; and a processor controlled by the stored programs for identifying the predicted classification of the at least one unclassified example wherein the processor includes: classification allocation means for allocating potential classifications to each unclassified example and for generating a plurality of classification sets, each classification set containing the plurality of training classified examples and the at least one unclassified example with its allocated potential classification; assay means for determining a strangeness value for each classification set; and a comparative device for selecting a classification set containing the most likely allocated potential classification for the at least one unclassified example, whereby the predicted classification output by the output terminal is the most likely allocated potential classification, according to the strangeness values assigned by the assay means.

In a third aspect the present invention provides a data classification method comprising:

inputting a plurality of training classified examples and at least one unclassified example;

identifying a predicted classification of the at least one unclassified example which includes

allocating potential classifications to each unclassified example;

generating a plurality of classification sets each containing the plurality of training classified examples and the at least one unclassified example with an allocated potential classification;

determining a strangeness value for each classification set; and

selecting, according to the assigned strangeness values, a classification set containing the most likely allocated potential classification; and

outputting the predicted classification for the at least one unclassified example whereby the predicted classification output by an output terminal is the most likely allocated potential classification.

An example of the present invention will now be described by way of example only with reference to the accompanying drawings, in which:

Figure 1 is a schematic diagram of data classification apparatus in accordance with the present invention;

Figure 2 is a schematic diagram of the operation of data classification apparatus of Figure 1;

Figure 3 is a table showing a set of training examples and unclassified examples for use with a data classifier in accordance with the present invention; and

Figure 4 is a tabulation of experimental results where a data classifier in accordance with the present invention was used in character recognition.

In Figure 1 a data classifier 10 is shown generally consisting of an input device 11, a processor 12, a memory 13, a ROM 14 containing a suite of programs accessible by the processor 12 and an output terminal 15. The input device 11 preferably includes a user interface 16 such as a keyboard or other conventional means for communicating with and inputting data to the processor 12 and the output terminal 15 may be in the form of a display monitor or other conventional means for displaying information to a user. The output terminal 15 preferably includes one or more output ports for connection to a printer or other network device. The data classifier 10 may be embodied in an Application Specific Integrated Circuit (ASIC) with additional RAM chips. Ideally, the ASIC would contain a fast RISC CPU with an appropriate Floating Point Unit.

To assist in an understanding of the operation of the data classifier 10 in providing a prediction of a classification for unclassified (unknown) examples, the following is an explanation of the mathematical theory underlying its operation.

Two sets of examples (data vectors) are given: the training set

consists of examples with their classifications (or *classes)* known and a test set consisting of unclassified examples. In Figure 3, a training set of five examples and two test examples are shown, where the unclassified examples are images of digits and the classification is either 1 or 7.

The notation for the size of the training set is *l* and, for simplicity, it is assumed that the test set of examples contains only one unclassified example. Let (X,A) be the measurable space of all possible unclassified examples (in the case of Figure 3, X might be the set of all 16 *x* 16 grey-scale images) and (Y,B) be the measurable space of classes (in the case of Figure 3, Y might be the 2-element set {1, 7}). Y is typically finite.

The confidence prediction procedure is a family $\{f_\beta : \beta \in (0, 1]\}$ of measurable mappings $f_\beta : (X \times Y)^l \times X \to B$ such that:

1. For any confidence level β (in data classification typically we are interested in β close to 1) and any probability distribution P in $X \times Y$, the probability that

$$y_{l+1} \in f_\beta(x_1, y_1, \ldots, x_l, y_l, x_{l+1})$$

is at least β, where $(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y_{l+1})$ are generated independently from P.

2. If $\beta_1 < \beta_2$, then, for all $(x_1, y_1, \ldots, x_l, y_l, x_{l+1}) \in (X \times Y)^l \times X$,

$$f_{\beta_1}(x_1, y_1, \ldots, x_l, y_l, x_{l+1}) \subseteq f_{\beta_2}(x_1, y_1, \ldots, x_l, y_l, x_{l+1})$$

The assertion implicit in the prediction $f_{\beta_1}(x_1, y_1, \ldots, x_l, y_l, x_{l+1})$ is that the true label y_{l+1} will belong to $f_{\beta_1}(x_1, y_1, \ldots, x_l, y_l, x_{l+1})$. Item 1 requires that the prediction given by f_β should be correct with probability at least β, and item 2 requires that the family $\{f_\beta\}$ should be consistent: if some label *y* for the (*l*+1)th example is allowed at confidence level β_1, it should also be allowed at any confidence level $\beta_2 > \beta_1$.

A typical mode of use of this definition is that some conventional value of β such as 95% or 99%, is chosen in advance, after which the function f_β is used for prediction. Ideally, the prediction region output by f_β will contain only one classification.

An important feature of the data classification apparatus is defining f_β in terms of solutions α_i, i=1, ..., l+1, to auxiliary optimisation problems of the kind outlined in US5640492, the contents of which is incorporated herein by reference. Specifically, we consider $|Y|$ completions of our data

$$(x_1, y_1),\ldots,(x_l, y_l), x_{l+1}$$

the completion y, $y \in Y$, is

$$(x_1, y_1),\ldots,(x_l, y_l),(x_{l+1}, y)$$

(so in all completions every example is classified).

With every completion

$$(x_1, y_1),\ldots,(x_l, y_l),(x_{l+1}, y_{l+1})$$

(for notational convenience we write y_{l+1} in place of y here) is associated the optimisation problem

$$\frac{1}{2}(w \cdot w) + C\left(\sum_{i=1}^{l+1} \xi_i\right) \longrightarrow \min \qquad (1)$$

(where C is a fixed positive constant)

subject to the constraints

$$y_i((x_i \cdot w) + b) \geq \xi_i, i = 1,\ldots,l+1 \qquad (2)$$

This problem involves non-negative variables $\xi_i \geq 0$, which are called *slack variables*. If the constant C is chosen too large, the accuracy of solution can become unacceptably poor; C should be chosen as large as possible in the range in which the numerical accuracy of solution remains reasonable. (When the data is linearly separable, it is even possible to set C to infinity, but since it is rarely if ever possible to tell in advance that all completions will be linearly separable, C should be taken large but finite.)

The optimisation problem is transformed, via the introduction of Lagrange multipliers α_i, i=1, ..., l+1, to the dual problem: find α_i from

$$\sum_{i=1}^{l+1} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l+1} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) \longrightarrow \max \qquad (3)$$

under the "box" constraints

$$0 \le \alpha_i \le C, \; i=1, 2, \ldots, l+1 \qquad (4)$$

The unclassified examples are represented, it is assumed, as the values taken by n numerical attributes and so $X = \mathbf{R}^n$.

This quadratic optimisation problem is applied not to the attribute vectors x_i themselves, but to their images $\mathbf{V}(x_i)$ under some predetermined function $\mathbf{V}: X \to \mathbf{H}$ taking values in a Hilbert space, which leads to replacing the dot product $x_i \cdot x_j$ in the optimisation problem (3)—(4) by the kernel function

$$\mathbf{K}(x_i, x_j) = \mathbf{V}(x_i) \cdot \mathbf{V}(x_j)$$

The final optimisation problem is, therefore,

$$\sum_{i=1}^{l+1} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l+1} y_i y_j \alpha_i \alpha_j \mathbf{K}(x_i, x_j) \to \max$$

under the "box" constraints

$$0 \le \alpha_i \le C, \; i=1, 2, \ldots, l+1$$

this quadratic optimisation problem can be solved using standard packages.

The Lagrange multiplier α_i, $i \in \{1, \ldots, l+1\}$, reflects the "strangeness" of the example (x_i, y_i); we expect that α_{l+1} will be large in the wrong completions.

For $y \in Y$, define

$$d(y) := \frac{\left| \{ i : \alpha_i \ge \alpha_{l+1} \} \right|}{l+1}$$

therefore $d(y)$ is the p-value associated with the completion y (y being an alternative notation for y_{l+1}). The confidence prediction function f, which is at the core of this invention, can be expressed as

$$f_\beta(x_1, y_1, \ldots, x_l, y_l, x_{l+1}) := \{ y : d(y) > 1 - \beta \}$$

The most interesting case is where the prediction set given by f_β is a singleton; therefore, the most important features of the confidence prediction procedure $\{f_\beta\}$ at the data $(x_1, y_1),...,(x_l, y_l), x_{l+1}$ are:

- the largest $\beta=\beta_0$ for which $f_\beta((x_1, y_1),...,(x_l, y_l), x_{l+1})$ is a singleton (assuming such a β exists);
- the classification $\mathbf{F}((x_1, y_1),...,(x_l, y_l), x_{l+1})$ defined to be that $y \in Y$ for which $f_{\beta_0}((x_1, y_1,...,(x_l, y_l), x_{l+1})$ is $\{y\}$.

$\mathbf{F}((x_1, y_1),...,(x_l, y_l), x_{l+1})$ defined in this way is called the f-optimal prediction algorithm; the corresponding β_0 is called the confidence level associated with **F**.

Another important feature of the confidence estimation function $\{f_\beta\}$ at the data $(x_1, y_1),...,(x_l, y_l), x_{l+1}$ is the largest $\beta=\beta_*$ for which $f_\beta((x_1, y_1),...,(x_l, y_l), x_{l+1})$ is the empty set. We call 1-β_* the credibility of the data set $(x_1, y_1),...,(x_l, y_l), x_{l+1}$; it is the p-value of a test for checking the iid assumption. Where the credibility is very small, either the training set $(x_1, y_1),...,(x_l, y_l)$ or the new unclassified example x_{l+1} are untypical, which renders the prediction unreliable unless the confidence level is much closer to 1, than is 1-β_*. In general, the sum of the confidence and credibility is between 1 and 2; the success of the prediction is measured by how close this sum is to 2.

With the data classifier of the present invention operated as described above, the following menus or choices may be offered to a user:

1. Prediction and Confidence
2. Credibility
3. Details.

A typical response to the user's selection of choice 1 might be prediction: 4, confidence: 99%, which means that 4 will be the prediction output by the f-optimal **F** and 99% is the confidence level of this prediction.

A typical response to choice 2 might be credibility: 100%, which gives the computed value of credibility. A typical response to choice 3 might be:

0	1	2	3	4	5	6	7	8	9
0.1%	1%	0.2%	0.4%	100%	1.1%	0.6%	0.2%	1%	1%

the complete set of p-values for all possible completions. The latter choice contains the information about $\mathbf{F}((x_1, y_1),...,(x_l, y_l), x_{l+1})$ (the character corresponding to the largest p-value), the confidence level (one minus the second largest p-value) and the credibility (the largest p-value).

This mode of using the confidence prediction function f is not the only possible mode: in principle it can be combined with any prediction algorithm. If **G** is a prediction algorithm, with its prediction $y:=\mathbf{G}((x_1, y_1),...,(x_l, y_l), x_{l+1})$ we can associate the following measure of confidence:

$$c(y) := \max\{\beta : f_\beta(x_1, y_1,..., x_l, y_l, x_{l+1}) \subseteq \{y\}\}$$

The prediction algorithm **F** described above is the one that optimises this measure of confidence.

The table shown in Figure 4 contains the results of an experiment in character recognition using the data classifier of the present invention. The table shows the results for a test set of size 10, using a training set of size 20 (not shown). The kernel used was $\boldsymbol{K}(x, y) = (x \cdot y)^3 / 256$.

It is contemplated that some modifications of the optimisation problem set out under equations (I) and (2) might have certain advantages, for example,

$$\frac{1}{2}(w \cdot w) + C\left(\sum_{i=1}^{l+1} \xi_i^2\right) \to \min,$$

subject to the constraints

$$y_i((x_i \cdot w) + b) = 1 - \xi_i, i = 1,...,l+1$$

It is further contemplated that the data classifier described above may be particularly useful for predicting the classification of more than one example simultaneously; the test statistic used for computing the p-values corresponding to different completions might be the sum of the ranks of αs

corresponding to the new examples (as in the Wilcoxon rank-sum test).

In practice, as shown in Figure 2, a training dataset is input 20 to the data classifier. The training dataset consists of a plurality of data vectors each of which has an associated known classification allocated from a set of classifications. For example, in numerical character recognition, the set of classifications might be the numerical series 0—9. The set of classifications may separately be input 21 to the data classifier or may be stored in the ROM 14. In addition, some constructive representation of the measurable space of the data vectors may be input 22 to the data classifier or again may be stored in the ROM 14. For example, in the case of numerical character recognition the measurable space might consist of 16x16 pixellated grey-scale images. Where the measurable space is already stored in the ROM 14 of the data classifier, the interface 16 may include input means (not shown) to enable a user to input adjustments for the stored measurable space. For example, greater definition of an image may be required in which case the pixellation of the measurable space could be increased.

One or more data vectors for which no classification is known are also input 23 into the data classifier. The training dataset and the unclassified data vectors along with any additional information input by the user are then fed from the input device 11 to the processor 12.

Firstly, each one of the one or more unclassified data vectors is provisionally individually allocated 24 a classification from the set of classifications. An individual strangeness value α_i is then determined 25 for each of the data vectors in the training set and for each of the unclassified data vectors for which a provisional classification allocation has been made. A classification set is thus generated containing each of the data vectors in the training set and the one or more unclassified data vectors with their allocated provision classifications and the individual strangeness values α_i for each data vector. A plurality of such classification sets is then generated with the allocated provisional classifications of the unclassified data vectors being different for each classification set.

Computation of a single strangeness value, the p-value, for each classification set containing the complete set of training data vectors and unclassified vectors with their current allocated classification is then performed 26, on the basis of the individual strangeness values α_i determined in the previous step. This p-value and the associated set of classifications is transferred to the memory 13 for future comparison whilst each of the one or more unclassified data vectors is provisionally individually allocated with the same or a different classification. The steps of calculating individual strangeness values 25 and the determination of a p-value 26 are repeated in each iteration for the complete set of training data vectors and the unclassified data vectors, using different classification allocations for the unclassified data vectors each time. This results in a series of p-values being stored in the memory 13 each representing the strangeness of the complete set of data vectors with respect to unique classification allocations for the one or more unclassified data vectors.

The p-values stored in the memory are then compared 27 to identify the maximum p-value and the next largest p-value. Finally, the classification set of data vectors having the maximum p-value is supplied 28 to the output terminal 15. The data supplied to the output terminal may consist solely of the classification(s) allocated to the unclassified data vector(s), which now represents the predicted classification, from the classification set of data vectors having the maximum p-value.

Furthermore, a confidence value for the predicted classification is generated 29. The confidence value is determined based on the subtraction of the next largest p-value from 1. Hence, if the next largest p-value is large, the confidence of the predicted classification is small and if the next largest p-value is small, the confidence value is large. Choice 1 referred to earlier, provides a user with predicted classifications for the one or more unknown data vectors and the confidence value.

Where an alternative prediction algorithm is to be used, the confidence value will be computed by subtracting from 1 the largest p-value for the sets of training data vectors and new vectors classified differently

from the predicted (by the alternative method) classification.

Additional information in the form of the p-values for each of the sets of data vectors with respect to the individual allocated classifications may also be supplied (choice 3) or simply the p-value for the predicted classification (choice 2).

With the data classifier and method of data classification described above, a universal measure of the confidence in any predicted classification of one or more unknown data vectors is provided. Moreover, at no point is a general rule or multidimensional hyperplane extracted from the training set of data vectors. Instead, the data vectors are used directly to calculate the strangeness of a provisionally allocated classification(s) for one or more unknown data vectors.

While the data classification apparatus and method have been particularly shown and described with reference to the above preferred embodiment, it will be understood by those skilled in the art that various modifications in form and detail may be made therein without departing from the scope and spirit of the invention. Accordingly, modifications such as those suggested above, but not limited thereto, are to be considered within the scope of the invention.

CLAIMS

1. Data classification apparatus comprising:

an input device for receiving a plurality of training classified examples and at least one unclassified example;

a memory for storing the classified and unclassified examples;

an output terminal for outputting a predicted classification for the at least one unclassified example; and

a processor for identifying the predicted classification of the at least one unclassified example

wherein the processor includes:

classification allocation means for allocating potential classifications to each unclassified example and for generating a plurality of classification sets, each classification set containing the plurality of training classified examples and the at least one unclassified example with its allocated potential classification;

assay means for determining a strangeness value for each classification set;

a comparative device for selecting a classification set containing the most likely allocated potential classification for the at least one unclassified example, wherein the predicted classification output by the output terminal is the most likely allocated potential classification according to the strangeness values assigned by the assay means; and

a strength of prediction monitoring device for determining a confidence value for the predicted classification on the basis of the strangeness value of a classification set containing the second most likely allocated potential classification of the at least one unclassified example.

2. Data classification apparatus as claimed in claim 1, wherein the processor further includes an example valuation device which determines individual strangeness values for each training classified example and the at least one unclassified example having an allocated potential

classification.

3. Data classification apparatus as claimed in claim 2, wherein Lagrange multipliers are used to determine the individual strangeness values.

4. Data classification apparatus as claimed in either of claims 2 or 3, wherein the assay means determines a strangeness value for each classification set in dependence on the individual strangeness values of each example.

5. Data classification apparatus comprising:

an input device for receiving a plurality of training classified examples and at least one unclassified example;

a memory for storing the classified and unclassified examples;

stored programs including an example classification program;

an output terminal for outputting a predicted classification for the at least one unclassified example; and

a processor controlled by the stored programs for identifying the predicted classification of the at least one unclassified example

wherein the processor includes:

classification allocation means for allocating potential classifications to each unclassified example and for generating a plurality of classification sets, each classification set containing the plurality of training classified examples and the at least one unclassified example with its allocated potential classification;

assay means for determining a strangeness value for each classification set;

a comparative device for selecting a classification set containing the most likely allocated potential classification for the at least one unclassified example, wherein the predicted classification output by the output terminal is the most likely allocated potential classification according

to the strangeness values assigned by the assay means and

a strength of prediction monitoring device for determining a confidence value for the predicted classification on the basis of the strangeness value of the classification set containing the second most likely allocated potential classification of the at least one unclassified example.

6. A data classification method comprising:

inputting a plurality of training classified examples and at least one unclassified example;

identifying a predicted classification of the at least one unclassified example which includes,

allocating potential classifications to each unclassified example;

generating a plurality of classification sets, each classification set containing the plurality of training classified examples and the at least one unclassified example with its allocated potential classification;

determining a strangeness value for each classification set;

selecting a classification set containing the most likely allocated potential classification for the at least one unclassified example wherein the predicted classification is the most likely allocated potential classification in dependence on the strangeness values;

determining a confidence value for the predicted classification on the basis of the strangeness value of the classification set containing the second most likely allocated potential classification for the at least one unclassified example; and

outputting the predicted classification for the at least one unclassified example and the confidence value for the predicted classification.

7. A data classification method as claimed in claim 6, further including determining individual strangeness values for each training classified example and the at least one unclassified example having an allocated

potential classification.

8. A data classification method as claimed in any one of the preceding claims, wherein the selected classification set is selected without the application of any general rules determined from the training set.

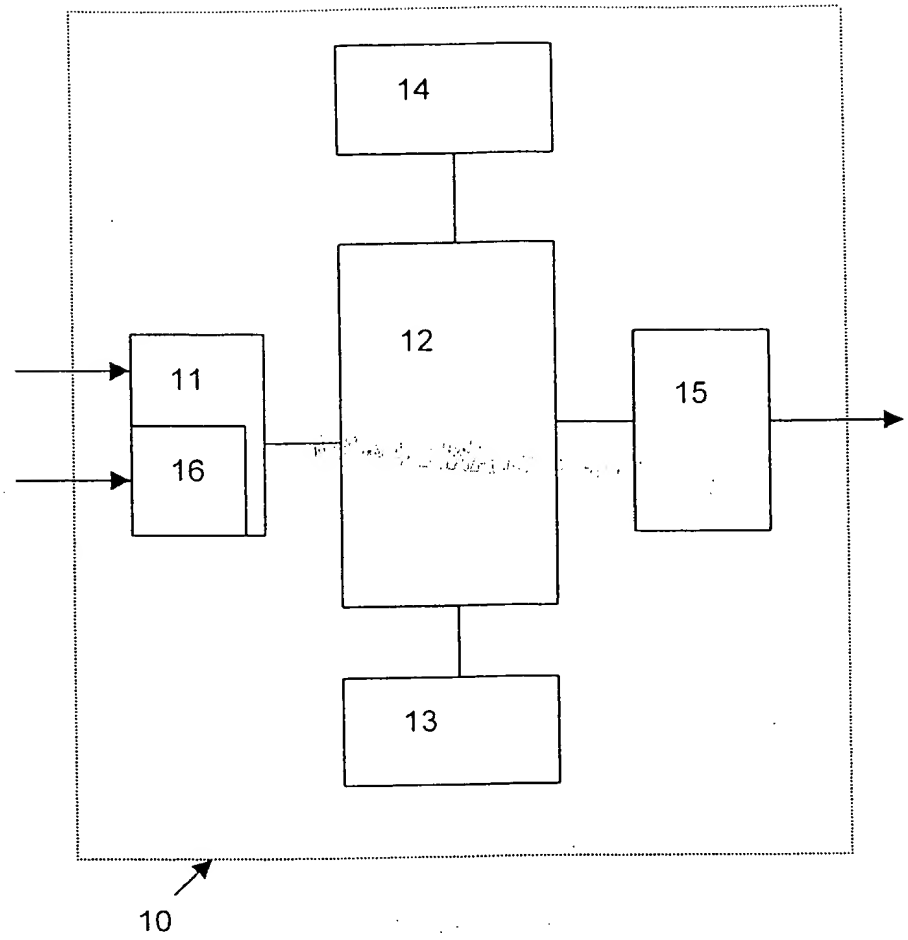

Figure 1

	Training Set					Test Set	
Example No.	1	2	3	4	5	1	2
Example	1	7	1	7	7	7	7
Classification	1	7	1	7	7	?	?

Figure 3

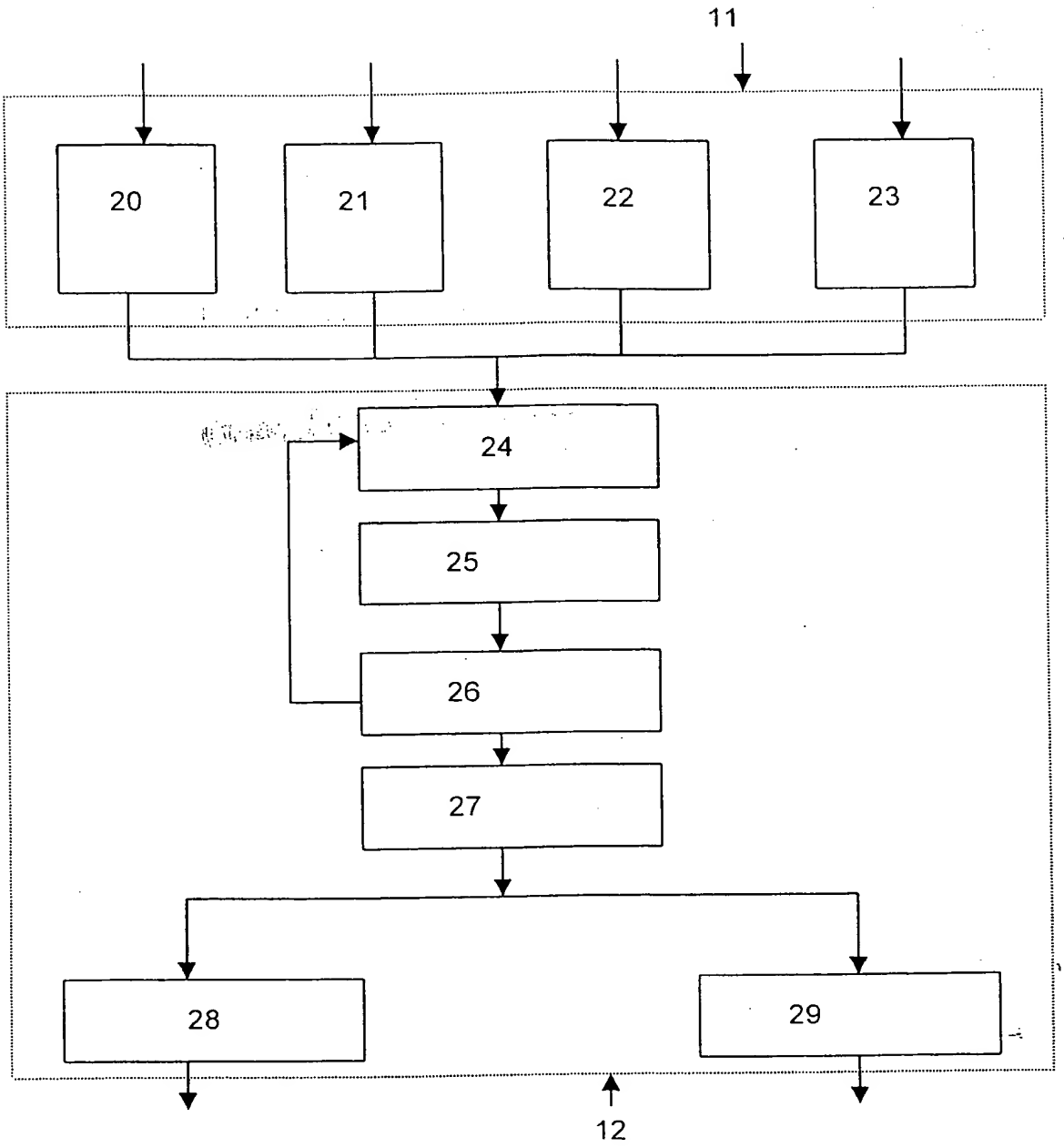

Figure 2

	Test Set									
Example No.	1	2	3	4	5	6	7	8	9	10
Example	1	7	7	1	7	1	1	1	7	7
True Class	1	7	7	1	7	1	1	1	7	7
Predicted Class	1	7	7	1	7	1	1	1	7	7
Confidence	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%
Credibility	19%	100%	100%	100%	100%	28%	100%	100%	100%	100%

Figure 4

PCT/GB99/03515

9. Nov. 99

Stevens Hewlett & Perkins